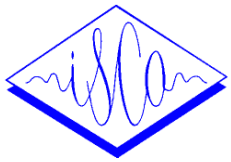

7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

EXPERIMENTS IN CONFIDENCE SCORING FOR WORD AND SENTENCE VERIFICATION

M. Andorno ⋆ and P. Laface ⋆ and R. Gemello ◇

⋆ Politecnico di Torino - Corso Duca degli Abruzzi, 24 - I-10129 Torino, Italy
E-Mail andorno/laface@polito.it
◇ Loquendo SpA - Via Nole, 55 - I-10149 Torino, Italy
E-Mail Roberto.Gemello@loquendo.com

ABSTRACT

The successful deployment of a telephone speech application cannot only rely on the accuracy of the recognition results, but also on their reliability. Reliable confidence measures are, thus, necessary in all practical applications to decide whether a recognized word - or sentence - should be accepted or rejected. Since most of the applications are based on continuous speech recognition, controlled by grammars, we present the results of a set of experiments aiming at assessing the quality and the limitations of different confidence measures for six different grammars that can be embedded in several applications. We show that using application independent confidence scoring techniques, good performance are obtained across all six grammars.
We introduce also a sentence level confidence measure that allows a significant reduction of the system error rate due to ill-formed sentences.

1. INTRODUCTION

It is well known that a reliable confidence measure associated to each hypothesis produced by a speech recognizer is an information of relevant value. It can be exploited in several different frameworks and applications: for example in Out-Of-Vocabulary (OOV) word detection, for keyword spotting, for unsupervised training, for reordering the hypotheses in an N-best decoder, or even during the decoding process.
The confidence measures are used in most telephone applications to allow the dialog system to rely on the (parts of) sentences that have been reliably detected. These applications often make use of continuous speech recognition, controlled by grammars of different complexity, for carrying out their task.
In this paper we present the results of a set of experiments aiming at assessing the quality and the limitations of different confidence measures for six different grammars that can be embedded in several applications. We are mainly interested in application independent confidence measures and scoring techniques that are fast to compute, that do not require held out data for training or modification of our recognizer architecture.
We show that combining an acoustic based confidence measure and a weighted N-Best score for tagging misrecognized, but in vocabulary words, good performance are obtained across all six grammars. We also propose a sentence level acoustic likelihood ratio measure to detect ill-formed sentences that do not include OOV, and that cannot be easily rejected using word confidence measures only.

This work was partially supported by the EU ITS Project SMADA - Speech Driven Multi-modal Automatic Directory Assistance

2. CONFIDENCE MEASURES

The Loquendo-ASR decoder uses a hybrid HMM-NN model where each phonetic unit is described in terms of a single or double state left-to-right automaton with self-loops. The Italian model is based on a set of 391 vocabulary and gender independent units including stationary context-independent phonemes and diphone-transition coarticulation models. The posterior probability of each state of the units given an acoustic feature vector is estimated by a Multilayer Perceptron. The HMM transition probabilities are uniform and fixed [1].
The confidence measures used in this work are based only on the posterior probabilities generated by the decoder or estimated from the N-best lists. We have also performed preliminary experiments to compare the confidence measures derived from the N-best lists with the ones derived from the word lattice [7].

2.1. Posterior probability based confidence measures

Confidence measures based on local phone posterior probability estimates generated by a hybrid HMM/NN model have been proposed in [9, 2]. To account for the raw acoustic information associated to each frame, the best score has been proposed as a measure of the matching between the data and the model [3]. In this approach, each utterance frame is scored against every output distribution in their HMMs to find the best score, independent of any information given by the sequence of phonetic units or words.
Building on these ideas, we propose as a word confidence measure the Acoustic Log Likelihood Ratio defined as:

$$ALLR(w_i) = \frac{\sum_{n=b}^{e} \max_{s \in S} \log P(s|o_n)}{\sum_{n=b}^{e} \log P(s^*_{w_i}|o_n)} \qquad (1)$$

where w_i is a word, b and e its beginning and ending frames according to the Viterbi segmentation, S is the number of output states of the NN model, o_n is the $n-th$ acoustic observation vector, and $s^*_{w_i}$ is the sequence of states produced by the Viterbi alignment of

the sequence of acoustic vectors o_b^e against the w_i HMM.
$ALLR(w_i)$ is, thus, the ratio between the free score, given by the sum of the a posteriori log probability of the best matching state for each frame, and the sum of the frame scores constrained by the word w_i model. This measure is easily computed in a hybrid HMM/NN model because all the probabilities are computed in parallel for each frame. The $ALLR(w_i)$ values range from 0 to 1, and the maximum is reached when the free score and the constrained one for each frame are the same, indicating an optimal acoustic matching according to the model. Low values of $ALLR(w_i)$ are, instead, good indicators of acoustic mismatch and of OOV words.

2.2. N-best based confidence measures

A Weighted N-Best (WNB) stability is a commonly applied measure [6]. It is defined as the ratio between the sum of the utterance likelihoods for all the hypotheses including a given word, and the sum of all the likelihoods in the N-best list.
However, considering the grammar of connected digits, where the same digit may appear several time in an utterance, we decided to account only for the sentence hypotheses in which a word appears approximately in the same time frames. In particular, we consider that two word hypotheses refer to the same word if their overlap region is at least 50% of their duration.
The overhead for computing the confidence measure for all the words of the best hypothesis is minimal since it is carried out within the module that produces the N-best hypotheses.

2.3. Product of the ALLR and WNB confidence measures

The third confidence measure that we have tested is the product of the acoustic and of the N-best based confidence measures

$$prod(w_i) = ALLR(w_i) * WNB(w_i)^{\alpha} \quad (2)$$

where the parameter α has been set once for all and not optimized for each testset. Other linear combinations did not give clearly better results.

3. EVALUATION OF CONFIDENCE MEASURES

Several evaluation metrics for confidence measures have been proposed. We will present our results by means of the Detection Error Tradeoff curves, but also through the single performance value of the Normalized Cross Entropy [3, 4].
Another interesting curve can be derived from the measure of the False Acceptance rate at a False Rejection rate of x%, referred to in the following as $FA@x\%FR$.
Using $FA@x\%FR$ and the baseline error rate p_e, two other values can be obtained: the rejection rate

$$Rej(@x\%FR) = x * (1 - p_e) + (1 - FA@x\%FR) * p_e \quad (3)$$

and the error rate of the recognizer on the accepted words

$$Err(@x\%FR) = \frac{FA@x\%FR * p_e}{(1 - x) * (1 - p_e) + FA@x\%FR * p_e} \quad (4)$$

The Rejection Error Tradeoff curve $(Rej(\tau), Err(\tau))_{\tau=0}^{1}$, summarizes the tradeoff between the rejected hypotheses (both correct and incorrect) and the error rate on the remaining hypotheses in the testset. The curve will be plotted showing in the y-axis the relative error rate reduction percent, $100 * (1.0 - Err(\tau)$.

Grammar	Nodes	Arcs	|Voc.|	Sentences	Words
Digits	4	9	11	860	13412
Phone No.	5	13	999	2889	24493
Integer $[0 - 10^9[$	65	160	117	2189	4907
Time of the day	61	268	125	2430	6758
Date	73	314	145	2827	10400
Euro currency	183	637	125	2604	8673
Loop	No LM		9400	1000	6916

Table 1: Test sets. Column |Voc.| shows the maximum size of the vocabulary associated to a grammar node.

4. EXPERIMENTS

4.1. Test sets and grammars

The six test sets and grammars that have been used for experiments described in this paper are reported in Table 1. The complexity of the grammars, well correlated with the baseline word error rates, increases from top to bottom.
The first set of experiments has been performed on continuous speech recognition *without* language modeling on a subset of 1000 sentences of the SpeechDat2 database, with a vocabulary of 9400 words as shown in the last row of Table 1. The aim was to determine if a lattice based method [7] clearly outperform the Weighted N-Best approach.
Table 2 shows the results, in terms of Normalized Cross Entropy, for the Acoustic Log Likelihood Ratio, Weighted N-Best and lattice based confidence measures. As expected the ALLR measure is very poor for in vocabulary words of a large vocabulary, while the Weighted N-Best and lattice based confidence measures are almost equivalent. Since the WNB confidence is easier and faster to be computed, the lattice based approach was not tested in remaining experiments.
Table 3 shows, for each grammar, and for the In Grammar and Out Of Grammar testsets, the Normalized Cross Entropy obtained using the ALLR, Weighted N-Best, and the $prod(w_i)$ confidence measures. The WNB confidence performs better than the ALLR for the well-formed grammar sentences, while the reverse is true if a set of out of vocabulary/grammar sentences is added to each testset. The $prod(w_i)$ of the two confidence measures approaches the behavior of the best one in the two cases.
In order to evaluate the confidence measures in a condition closer to a real application, we added to the In Grammar testsets 5% of ill-formed utterances of the same domain for the Time, Date and Euro grammars.
Table 4 reports the Error rate reduction and total Rejections that are obtaining by setting the threshold on the $prod(w_i)$ measure to a value that gives a False Rejection rate of 5%, quite interesting from an application point of view.

Acoustic LLR	Weighted N-Best	Lattice based
0.052	0.218	0.210

Table 2: NCE for three confidence measures on the SpeechDat2 testset

	NCE for In Vocabulary/Grammar					
CM	Digits	Phone	Integer	Time	Date	Euro
ALLR	0.287	0.170	0.126	0.117	0.251	0.134
WNB	0.370	0.279	0.372	0.332	0.441	0.217
Prod	0.382	0.280	0.373	0.317	0.438	0.282
	NCE for In + Out of Vocabulary/Grammar					
CM	Digits	Phone	Integer	Time	Date	Euro
ALLR	0.864	0.690	0.646	0.505	0.656	0.536
WNB	0.348	0.274	0.196	0.319	0.369	0.153
Prod	0.843	0.565	0.623	0.571	0.674	0.521

Table 3: Normalized Cross Entropy using the ALLR, Weighted N-Best, and the $prod(w_i)$ confidence measures

	In Vocabulary/Grammar					
(%)	Digits	Phone	Integer	Time	Date	Euro
Err red.	80.0	50.3	69.5	55.5	75.9	49.9
Rej	5.4	5.9	6.4	7.0	7.3	7.2
	In + Quasi In Vocabulary/Grammar					
Err red.				50.0	69.9	48.4
Rej				8.7	8.1	8.4

Table 4: In vocabulary/grammar sentences: Error rate reduction and total Rejections setting a threshold on the $prod(w_i)$ measure that gives a False Rejection of 5%.

The original error, for the in vocabulary/grammar sentences, is reduced in most of the cases by more than 50% accepting a total number of rejections less than 7.3%. Including in the testset the ill-formed utterances, the error reduction obtained by the $prod(w_i)$ confidence measure remains still at the same level while the total Rejections percentage slightly increases. The behavior of the three confidence measures can be also compared in the DET curves of figure 1.

5. UTTERANCE VERIFICATION

5.1. Rejection of out of grammar utterances

For the rejection of out of grammar utterances, the $ALLR(w_i)$ confidence measures are combined in different ways to obtain a confidence measure at the sentence level [5]. In particular, figure 2 shows the DET curves for a set of sentences including 25% out of vocabulary/grammar sentences for the Date grammar. In our experiments, for all the grammars, the best combination of the word level confidence measures for detecting out of grammar sentences is $meanALLR(w_i)$, the average of the confidence scores of the words in the sentence, while $minALLR(w_i)$ performs better for the "quasi well-formed" utterances described in the next section.

5.2. Rejection of "quasi well-formed" utterances

The rejection of quasi in grammar utterances is very difficult using the measures based on the posterior probabilities and those estimated from the N-best lists introduced in the previous sections. The $ALLR$ measures are useful for the OOV word detection, but "quasi well-formed" utterances very often do not include OOV words. The Weighted N-Best stability measure does not help either, because In Vocabulary words - or misrecognized words - within an ill-formed sentence may induce to errors the grammar constrained decoding, while obtaining high confidence scores. The possibility of accepting ill-formed sentences increases when the grammar includes similar vocabularies in many of its nodes; this is the case of the Date grammar.

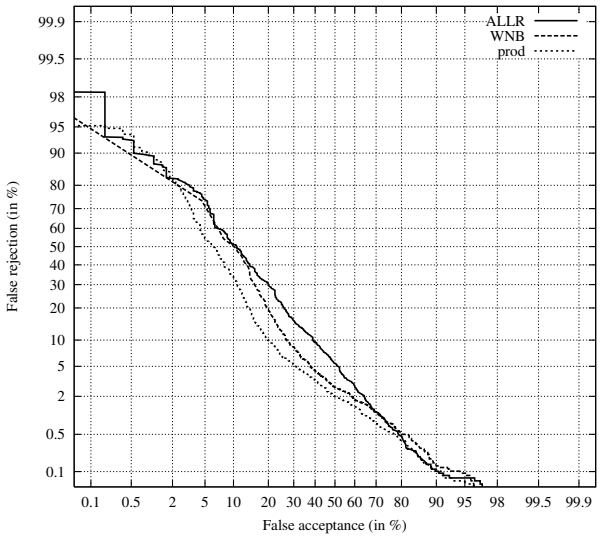

Figure 1: Date grammar: Detection Error Tradeoff curve for a testset including 5% "Quasi In Grammar" utterances.

Let's consider, for example, the sentence "Sedici Aprile Duemila" (16th of April 2000), that is well-formed for the Date grammar. Suppose the end-point detector triggers the end of sentence just after the word "sedici" has been pronounced. Since "sedici" is an ill-formed sentence for the Date grammar, while the acoustically similar sequence "sei dieci" (6th of October) is in grammar, it is likely that the grammar constrained decoding produces several N-best hypotheses including the sentence "sei dieci".

Our approach to this problem is to merge in a single looped grammatical node the union of all the vocabularies that are associated to the grammar nodes. Then we perform a recognition step to obtain a free-of-grammar likelihood for the input utterance. A Sentence Log Likelihood Ratio ($SLLR$) confidence measure is then defined, in analogy with the $ALLR$, as the ratio of the free-of-grammar likelihood and the grammar constrained score.

In the previous example, the free-of-grammar likelihood for "sedici" would be better than the one of the grammar constrained one "sei dieci", reducing the $SLLR$.

The free-of-grammar recognition step is much faster than the grammar constrained one because the grammar is simpler and the state output probabilities are already available.

Figure 3 shows the Rejection Error Rate Tradeoff for a testset including 5% of ill-formed but in vocabulary sentences, where a dramatic error rate reduction can be observed in comparison with the best competitor $minALLR(w_i)$ confidence measure. By rejecting 6% of the sentences, the error rate for the remaining ones decreases by more than 70%. Similar results were obtained for the other grammars.

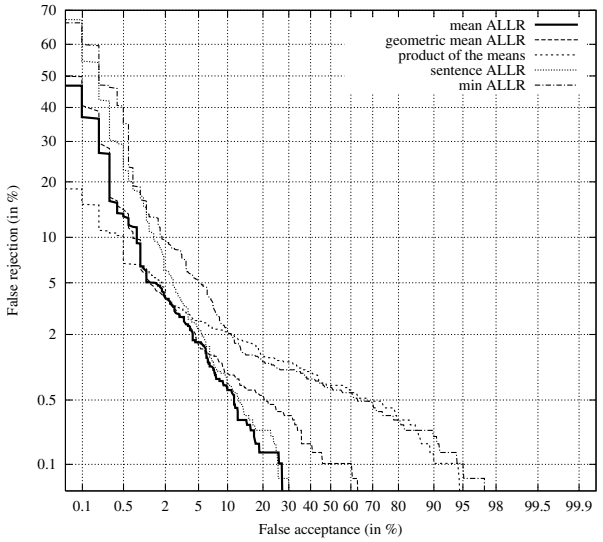

Figure 2: Date grammar: DET for a testset including 25% out of vocabulary sentences.

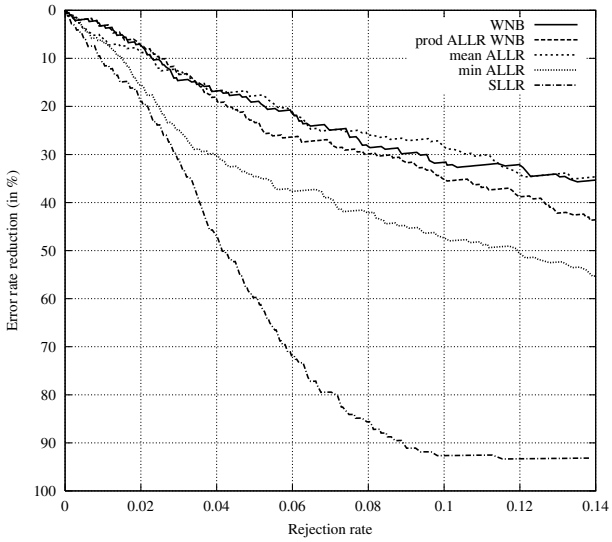

Figure 3: Date grammar: Rejection Error Rate Tradeoff for a testset including 5% of ill-formed but in vocabulary sentences.

A rejection strategy that account for the results of these experiments has been tested by adding to the Date testset both the 5% ill-formed sentences and the 25% out of vocabulary/grammar sentences. The rejection procedure is as follows:

1. reject sentences with a $meanALLR(w_i)$ value less than 0.5 (sentences including OOV)
2. reject sentences with a $SLLR$ value less than 0.95 (ill-formed sentences)
3. process the remaining (putative well-formed) sentences according to the $prod(w_i)$ confidence measure.

Table 5 shows the percentage of sentences rejected for each category after the first and second step. The DET curve on the sentences produced by the final step closely matches the one of the Date In Grammar testset, confirming that almost all the sentences that were not rejected belong to the in grammar category.

Step	In grammar	Quasi In Grammar	Out of Grammar
1	1%	10%	92%
2	5%	90%	96%

Table 5: Percentage of sentences rejected for each category

6. CONCLUSIONS

We introduced a rejection procedure that uses application independent confidence scoring techniques, and allows a significant reduction of the system error rate due to out of grammar of ill-formed sentences. Good performance were obtained for six grammars using a simple $prod(w_i)$ combination of word confidence scores for tagging misrecognized, but in vocabulary words.

Acknowledgments
We thank Daniele Colibro, Luciano Fissore, Franco Mana, and Claudio Vair for making available the grammars and test sets, and for helpful discussions.

7. REFERENCES

[1] D. Albesano, R. Gemello, and F. Mana, “Hybrid HMM-NN Modeling of Stationary-Transitional Units for Continuous Speech Recognition”, Int. Conf. On Neural Information Processing, pp. 1112–1115, 1997.

[2] G. Bernardis, and H. Bourlard, “Improving Posterior Based Confidence Measures in Hybrid HMM/ANN Speech Recognition Systems", Proc. ICSLP 9198, Sydney, pp. 775–778, 1998.

[3] L. Gillick, and Y. Ito, and J. Young, “A Probabilistic Approach to Confidence Estimation and Evaluation"’ Proc. ICASSP 1997, pp. 879–882, 1997.

[4] B. Maison, and R. Gopinath, “Robust Confidence Annotation and Rejection for Continuous Speech Recognition", Proc. ICASSP 01, 2001.

[5] B. Souvignier, and A. Wendemuth, “Combination of Confidence Measures for Phrases", Proc. ASRU 1999 Workshop, Keystone, USA, pp. 217-220, 1999.

[6] M. Weintraub, “LVCSR Log-Likelihood Ratio Scoring for Keyword Spotting", Proc. ICASSP 1995, pp.297–300, 1995.

[7] F. Wessel, K. Macherey, H. Ney, “A Comparison of Word Graph and N-Best List Based Confidence Measures", Proc. EUROSPEECH 1999, pp. 315–318, 1999.

[8] F. Wessel, K. Macherey, H. Ney, “Confidence Measures for Large Vocabulary Continuous Speech Recognition", IEEE Transactions on Speech and Audio Processing, Vol. 9, n. 3, pp. 288–298, March 2001

[9] G. Williams, and S. Renals, “Confidence Measures from Local Posterior Probability Estimates", Computer Speech and Language, Vol. 13, pp. 395–411, 1999.